DOCUMENT RESUME

ED 362 474 SP 034 676

AUTHOR Hummel, Thomas J.
TITLE A Research Simulation for Counselor Trainees.
PUB DATE Apr 93
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Descriptive (141)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Computer Simulation; Computer Software; Efficiency; Elementary Secondary Education; Graduate Study; Higher Education; *Instructional Effectiveness; Professional Education; *Research Papers (Students); *Research Skills; *School Counselors; School Psychologists; Skill Development; *Student Research
IDENTIFIERS University of Minnesota

ABSTRACT

The Counseling and Student Personnel Psychology program at the University of Minnesota sought to create a graduate-level research experience which would be both instructionally efficient and educationally effective. Important skills to be learned in doing research were identified, including the development of a research question, critical review of research related to the question, design of a research project, data analysis, presentation of results, and drawing appropriate conclusions. Data collection in itself was not felt to be a critical skill. A research simulation which included all critical components was developed. However, a flexible data source was needed which could tailor data to a student's unique research design. Using a data specification form, each student defined the population that would have been sampled had the research been carried out. The student then entered the form into the computer, and an artificial data set was generated by a computer program called "DataBook." Advantages of the simulation over a traditional research experience include: (1) students are required to focus on their variables in a very detailed manner; (2) the quality of statistical analysis is improved; (3) the simulation takes far less time than an actual research project; and (4) it simplifies taking a group of students through the experience together. Appendixes contain a student orientation handout, course syllabus, and a DataBook description and manual. (JDD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 362 474

A Research Simulation for Counselor Trainees

By

Thomas J. Hummel
University of Minnesota

A Paper Presented at the Annual Meeting of the
American Educational Research Association
Atlanta, 1993

(© Copyright 1993 by Thomas J. Hummel, All Rights Reserved)

SP 034 674

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

T. J. Hummel

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

The Counseling and Student Personnel Psychology (CSPP) program at the University of Minnesota faces pressures which are becoming common in the current economic environment. On the one hand, both our faculty and our major accrediting agency, the American Psychological Association, are concerned that our doctoral program be an intense, closely supervised training experience. To maintain the level of supervision required, however, implies that we must reduce the number of doctoral level students as we lose faculty in our program as part of a general downsizing of our college. The limitation on the number of doctoral students, in turn, reduces the number of student credit hours our program generates, contrary to the demands from our college for ever higher levels of instructional efficiency. We are faced with a need to decrease the size of our doctoral program and to increase the production of student credit hours at a time when our faculty is getting smaller and teaching loads were at an all time high.

A solution to these conflicting demands was to increase the size of our master's degree program in a way that would maintain, and perhaps even improve, the quality of the educational experience. A review of our options showed that the quality of instruction for many of the master's level courses would not be greatly affected by doubling or even tripling enrollments. This was not, however, true of the practicum or the research requirement. This presentation deals with the effort to create a research experience which would be both instructionally efficient and educationally effective.

Originally, the CSPP program required each student to complete a nine credit "Plan B" research project, which consisted of carrying out and writing up the student's data-based research. This research, which was similar in scope to a master's degree thesis, was supervised and evaluated by the student's advisor. It required a great deal of time from both faculty and students, much more, in fact, than required by the Graduate School for a "Plan B" master's degree. The faculty decided to develop a new research experience which would require fewer resources, but would teach more in the way of research skills. Central to this new experience would be a simulation that would teach the skills usually learned while carrying out research for a master's thesis.

Simulations have long been used to teach complex skills. Perhaps the most well known of these are the elaborate flight simulators used to train commercial and military pilots. Consistent with the need stated above, simulations are used to save resources by finding an economical way to learn and practice the critical components of a complex skill.

For master's level counselor trainees, important skills to be learned in doing research include the development of a research question, the critical review of research related to this question, the design of a research project, the analysis of data, the presentation of results, and the drawing of appropriate conclusions. Absent from this list is the actual collection of data. Through field experiences and practica, counselors are taught to work with people and to collect and organize information. Unlike students in many other disciplines, counselors administer tests and keep track of the results, i.e., they collect data. It seems reasonable, therefore, to suggest that data collection in itself is not a critical skill for counselors to learn during their research experience.

A research simulation which includes all components except data collection can take far less time than an actual research project, and it simplifies taking a group of students through the experience together. However, to be realistic, there must be a flexible source of data, one which can tailor a set of data to a student's unique research design, regardless of the number of groups or the kind and number of variables.

The faculty decided to embed the research simulation in a course designed to include both a traditional and a nontraditional component. The traditional part includes the use of a good, well structured research design text to broaden students' background in research methods and to teach them the "language of research."

The nontraditional part of the course is the research simulation. The simulation has the following major components: a research proposal, containing an introduction, critical review of literature and methods; a budget that allocates up to $300 and 300 hours of student effort; Data Specification Form(s); data acquisition; data analysis; and the presentation and interpretation of findings.

Rather than collecting data from subjects, the data are acquired from a computer. Using the Data Specification Form(s), each student defines the population(s) that would have been sampled had the research been carried out. The student then enters the Data Specification Form(s) into the computer, and his/her data set is produced. The data in hand, the student proceeds as he or she normally would.

The computer program which generates the artificial data sets is named *DataBook*. The output of the program is a data listing, one row for each "subject," including group identification, subject number, and the values of whatever variables were involved in the student's design. Primary design requirements for the program were that it be easy to use and provide realistic looking data. *DataBook* was developed using Asymetrix's *Toolbook* and Borland International's *Turbo Pascal for Windows*. *Toolbook* provided an efficient way to develop a graphical, "point and click" user interface, the kind now familiar to Windows and Macintosh users. *Turbo Pascal for Windows* was used to develop a dynamic link library (DLL) that contained the necessary numerical routines for generating the data.

With respect to the requirement for realism, it was important that the values of whatever variables were called for in the student's research design, e.g., MMPI scores, GPAs, high school ranks, or Likert scales, would look like those they would have recorded had they collected the data.

There are important similarities and differences between the research simulation the students complete and a traditional master's level research paper. The first three sections of the proposal are very similar to the first three chapters of a typical master's paper, except that section one is somewhat shorter and section two is considerably shorter. The analysis, tables presented, interpretation, and conclusions are the same as they would have

been in a thesis. There is, however, less speculative discussion and a less extensive recommendations section.

Whatever is learned by people who go out and collect their own data is not learned by these students as a part of this project. As stated above, the position here is that counselors learn to collect data in other ways. Further, some master's students in the past analyzed their advisor's or other's data and, therefore, were not exposed to the collection experience. It must be acknowledged, however, that some students learn some things from collecting data that the simulation cannot provide. Collecting your own data, the drudgery aside, can also provide, for some people, a sense of excitement and motivation at the time of analysis. That is hard to match with artificial data.

Given these advantages to actually collecting data, are there advantages for the simulation that balance these? The answer is "Yes." Filling out the Data Specification Forms requires students to focus on their variables in a very detailed manner. They use their imaginations and the literature they review to form hypotheses about population parameters. With respect to correlations between variables, they sometimes conjure up values that are mathematically impossible. *DataBook* catches these because it requires positive definite correlation matrices for multivariate data generation, and the student is sent back to the drawing board. Another advantage for simulation is quality of statistical analysis. Each student's data is on the computer in a form readable by statistical packages. Literally within seconds, the student's data can be loaded into a package and then analyzed. With a modern menu driven program, such as *SYSTAT,* Version 5.03, and given the level of complexity of most of the students' analyses, it takes only a minute or so to check the students work. In the past, this was impractical.

On balance, then, it is concluded that the research simulation is just as effective as a master's research paper for teaching the skills required by master's level counseling students. The two experiences have offsetting advantages in terms of student learning. In terms of cost, though, the simulation is a clear winner. This cannot be ignored in a time of shrinking resources.

In addition to cost and student learning, Dr. Rodney Loper, a clinical psychologist who frequently serves on CSPP students' committees, pointed out another benefit. He sees the simulation as a more ethical training approach. Students' initial research efforts are not inflicted upon subjects and the counseling agencies and student development offices where they are often found. Too many neophytes can be disruptive and make agencies less receptive to better trained researchers. Letting them sharpen their skills in a simulation allows their first "live" research experience to have a higher likelihood of being polished and professional.

Additional information on *DataBook* and the research simulation is included in this paper's appendices. Appendix A contains a student orientation handout that is distributed midway through the quarter preceding the simulation. This handout restates some of the above information, but it goes into more detail on the steps the students must carry out.

Appendix B contains the syllabus for the course in which the simulation is embedded. Included therein is a schedule that specifies when assignments are turned in, defining the time frame for the simulation. Included in Appendix C is the *DataBook* manual and a copy of the Data Specification Form. These items give a good idea of how to install and use the program. Finally, Appendix D provides a technical description of *DataBook*.

References

Moonan, W. J. Linear transformation to a set of stochastically dependent normal variables. *Journal of the American Statistical Association*, 1957, 52, 247-252.

Scheuer, E. M., & Stoller, D. S. On the generation of normal random vectors. *Technometrics*, 1962, 4, 278-281.

Zelen, M., & Severo, N. C. Probability functions. In M. Abramowitz & I. A. Stegun (Eds.), Handbook of Mathematical Function. New York: Dover, 1970, pp. 925-996.

Appendix A

Student Orientation Materials

CSPP Plan B Research Simulation (EPsy 8431 MA Research Seminar)

PURPOSE

The objective of this simulation is to teach you the skills usually learned while carrying out research for a master's thesis project.

BACKGROUND & RATIONALE

Simulations have long been used to teach complex skills. Perhaps the most well known of these are the elaborate flight simulators used to train commercial and military pilots. Simulations are used to learn and practice the critical components of a complex skill.

For CSPP master's level students, important skills to be learned in doing research are the development of a research question, the critical review of research related to this question, the design of a research project, the analysis of data, the presentation of results, and the drawing of appropriate conclusions. Absent from this list is the actual collection of data. Through your field experiences and practica, you are taught to work with people and to collect and organize information. Unlike students in many other disciplines, counseling students administer tests and keep track of the results, i.e., they collect data. Because of this, it seems reasonable to suggest that data collection is not a critical skill for counselors to learn during their research experience.

A research simulation which includes all components except data collection can take far less time than an actual research project, and it makes it possible to take a group of students through the experience together. However, to be realistic, there must be a flexible source of data, one that can tailor a set of data to your unique research design, regardless of the number of groups or the kind and number of variables.

DESCRIPTION

To develop an alternative to the traditional master's level research, EPsy 8431 was designed to have both a traditional and a nontraditional component. The traditional part includes the use of a good, well structured research design text to broaden students' background in research methods and to help them learn the "language of research."

The nontraditional part of the course is the research simulation. Central to this is the computer program named *DATABOOK*, which produces artificial data sets. The output

of the program is a data listing, one row for each "subject," including group identification, subject number, and the values of whatever variables are involved in your design.

DATABOOK is easy to use and provides realistic looking data. The values of whatever variables are called for in your research design, e.g., MMPI scores, GPAs, high school ranks, or Likert scales, look like those you would have recorded had they been collected.

COMPONENTS OF THE SIMULATION

The research simulation has the following major components:

1. Research proposal. The proposal contains three sections: introduction, critical review of literature, and methods. The introduction includes a statement of the problem and a subsection on the importance of the problem. The literature review aims for a critique of studies related to the problem rather than a simple summary. The methods section includes the usual elements concerning subjects, independent and dependent variables, procedures, hypotheses, and analysis, and varies depending on the problem under study. The three sections are turned in separately, critiqued, and returned to you for further development. After you have feedback on all sections, your proposal is resubmitted.

2. Budget. A written budget is required to keep your project realistic in terms of its size and scope. You are limited to $300 and 300 hours, which you can allocate to various research costs and activities. You can go beyond these limits only if you can make a sound argument as to need and availability of resources.

3. Data specification. You will complete a "Data Specification Form" for each independent group in your research design . With these forms you define the populations from which your data will be sampled. To do this, you need to develop hypotheses about means, standard deviations, distributions, and intercorrelations. You also need to determine the format of the data (e.g., whole numbers for MMPI scores, or decimal fractions for GPAs).

4. Data generation. You input your Data Specification Form(s) into the *DATABOOK* program, which in turn creates a copy of itself for you. In this way, you have your own "data definition book" which specifies the population(s) for your study. After you are finished, the instructor has the option of altering your data definition and thereby controlling the "true state" of affairs. The data definition complete, *DATABOOK* produces a listing of your data.

5. Data analysis. The analysis is accomplished in several ways. You may analyze your data on a personal computer with a statistical package you have, or you may use *MYSTAT*, which can be purchased through the bookstore for approximately $15. If

you design a study which calls for a statistical analysis with which you are not completely familiar, please consult with the instructor.

6. Data presentation. You will develop whatever tables and figures would typically be included in a master's thesis. For the most part, this entails presenting tables of descriptive and inferential statistics (e.g., ANOVA tables). These materials are turned in, critiqued by the instructor, returned, and revised.

7. Interpretation. Finally, you write up a summary of your findings, draw conclusions from your analysis, and make recommendations for future research. As with preceding sections, this one is turned in for feedback and returned.

At the end of the simulation, you assemble the revised sections and turn in your entire project.

EDUCATIONAL VALUE

There are important similarities and differences between your project and the traditional master's level research paper. The first three sections of the proposal are very similar to the first three chapters of a typical master's paper, except that section one of the proposal is somewhat shorter and section two is considerably shorter. The analysis, tables presented, interpretation, an conclusions are the same as would have been in a master's thesis. There is, however, less speculative discussion and less extensive recommendations.

Whatever is learned by people who go out and collect their own data is not learned as a part of this project. As stated above, the position here is that counselors learn to collect data in other ways. Further, some master's students in the past analyzed their advisor's or others' data and, therefore, were not exposed to the collection experience. It must be acknowledged, however, that some students learn some things from collecting data that the simulation cannot provide. Further, collecting your own data, the drudgery aside, can provide, for some people, a sense of excitement and motivation at the time of analysis. That is hard to match with artificial data.

Given that there are advantages to actually collecting data, are there advantages for the simulation that balance these? The answer is "Yes." Filling out data specification forms requires you to focus on your variables in a very detailed manner. You use your imagination and the literature you review to form hypotheses about population parameters. With respect to correlations between variables, you might conjure up values that are mathematically impossible. *DataBook* catches this because it requires positive definite correlation matrices for multivariate data generation. In this event, you return to the "drawing board" and develop more realistic values. Another advantage for simulation is quality of statistical analysis. Your data is on the computer in a form readable by statistical packages. Literally within seconds, your data can be loaded into a package and

analyzed. It takes only a minute or so to check your work. In the past this was impractical.

On balance, then, the research simulation is just as effective as a master's thesis for teaching the skills required by master's level counseling students. The two experiences have offsetting advantages in terms of student learning.

RULES OF THE GAME

1. Your research proposal must be written for EPsy 8431. You may not use a paper you have done for another class, e.g., a social psych class.

2. The topic must deal with counseling or student development, including psychological education.

3. The research must be quantitative in nature. Whatever methods you use, eventually, each "subject" will have one or more numbers associated with them.

4. Stay on schedule!

5. For this project, it is better to ask permission than to seek forgiveness. When in doubt, ask.

6. Finally, make sure the header, "Research Simulation," appears on each page you submit and include the following on the cover page of your finished project:

   This paper is submitted in partial fulfillment of the requirements for EPsy 8431 and the Plan B project. The results reported herein are based on simulated data.

Appendix B

Course Syllabus

EPsy 8431
Master's Seminar: CSPP
Spring, 1993

TEXT: Wiersma, W. *Research Methods in Education* (5th Ed.).

OBJECTIVES:

1. Know the various types of educational research and understand their associated research designs.

2. Know the sources of research literature and be able to integrate research reports.

3. Know the steps and procedures required to develop and carry out a research project, including:

   a. Developing the research problem
   b. Carrying out a literature review
   c. Designing the research approach
   d. Sampling populations
   e. Developing instruments
   f. Analyzing the data
   g. Writing the report

4. Appreciate the importance of ethics in research, including the integrity of the researcher and concerns for the rights of human subjects.

COURSE OUTLINE:

Week 1	Orientation to Educational Research - Types of educational research - Planning the research project - Legal and ethical constraints Wiersma, Ch. 1, 13
Week 2	Narrowing the Research Focus - Sources of research literature - Carrying out the literature review - Introduction to research critique Wiersma, Ch. 2, 3
Week 3	Critically Evaluating Research - Systematic literature analysis - Factors that affect quality of research - Effects related to the research situation

- Experimenter and statistical contamination
Wiersma, Ch. 4

Week 4 — Experimental Design
- Internal and external validity
- Random selection and random assignment
- Control-group designs
- Common mistakes in conducting experiments
Wiersma, Ch. 5

Week 5 — Quasi-Experimental and Correlational Research
- Studying the relationships between variables
- Using the causal-comparative method
- Using the correlational method
Wiersma, Ch. 6

Week 6 — MIDTERM
Measurement and Statistics

Week 7 — Measurement and Statistics, continued
Wiersma, Ch. 11, 12

Week 8 — Methods of Survey Research
- Sampling designs
- Survey research techniques and tools
Wiersma, Ch. 7, 10

Week 9 — Ethnographic Research
- Collecting data
- Observation and interviewing
- Analysis of data
Wiersma, Ch. 9

Week 10 — Historical Research
- Methodology of historical research
- Quantitative methods
Wiersma, Ch. 8

ASSIGNMENTS AND REQUIREMENTS:

A. Written Plan B Projects:

1. Research proposal. This written assignment is expected to be approximately 10 pages in length (min = 7, max = 12, double spaced, APA style).

   DUE DATES:

1st Part of Proposal (Introduction, 2-3 pages)	4/15/93
2nd Part of Proposal (Literature, 3-4 pages)	4/22/93
3rd Part of Proposal (Design, 2-3 pages)	4/29/93
Complete Proposal:	5/13/93

2. Data Specification Form(s) 4/29/93

3. Project Budget (Time, 300 hrs, & Money, $300) 4/29/93 (1 page)

4. Presentation of Results (Tables & Figures) 5/27/93

5. Interpretation, Conclusions, & Recommendations 5/27/93 (2-4 pages)

6. Complete Plan B project (Change tense in proposal, put tables and figures in the interpretation section, put Data Specification Form and budget in an appendix.) 6/8/93

B. Examinations:

1. Midterm Exam: 5/4/93

2. Final Exam: Tues. -- 6/8/93 -- 4:00-6:00 pm

Grading: To get a "C" or better, you must "pass" the examinations and the written projects. Having satisfactorily completed these, the specific grade will be assigned on the basis of 100 points (Midterm 20 pts., Final 30 pts., Written Project 50 pts.) The instructor may deduct up to 10 points for written assignments which are not handed in on time.

Appendix C

DataBook Manual and Data Specification Form

DataBook Installation and Use

System Requirements: Any MS-DOS system running Microsoft's *Windows 3.1* with 1.5 megabytes of free disk space.

Performance: Asymetrix's *Toolbook*, the system used to develop *Databook*'s graphical user interface, benefits from a fast computer and display adapter. Some *DataBook* "pages" have a complex screen layout and redraws will take some time on slower machines. On a 486DX50 machine, redraws take up to 3.5 seconds.

Preliminaries:

The following instructions are for installing the system using DOS commands. (Those familiar with Windows may wish to follow a similar set of steps using File Manager.) DOS names and screen output are in *capitalized italics*. Commands to be typed are in ***CAPITALIZED BOLD ITALICS***. After each command is typed, press the Enter key. Unless otherwise noted, it is assumed that after you finish typing anything you will press the Enter key. If the end of a command coincides with the end of a sentence in this text, a space is placed between the end of the command and the period. A similar convention is followed with respect to commas. No command in what follows contains a period or comma (.,). In the sections containing references to *Windows*, objects (e.g., windows, buttons, dialog boxes) appear in the **system font**. The term "click" has the usual *Windows* meaning, to press and release the left mouse button. Proprietary names, e.g., *Windows*, are italicized.

Installation:

1. Create a directory named *TOOLBOOK* under the root directory on your hard drive. To do this, first make sure you are on the "C drive" by typing ***C:*** , and then make sure you are in the root directory by typing ***CD *** . The last line on the screen should be *C:\>* . Now type ***MD TOOLBOOK*** .

2. Switch to the *TOOLBOOK* directory by typing ***CD TOOLBOOK*** . At this point, the last line on your screen should be *C:\TOOLBOOK>* .

3. Copy the files from the distribution disk, labeled "Runtime DataBook," to the *TOOLBOOK* directory. For example, if the distribution disk is in the "B drive," then type ***XCOPY B:*** .

4. To check on the installation, type ***DIR*** . The files in the left hand column of the following table should be listed in the directory.

Distribution files:

File	Size	Purpose
DATABOOK.TBK	87968	Main application file; defines user interface.
DATABOOK.DLL	23552	Dynamic link library (DLL); statistical computations
TBOOK.EXE	396320	Runtime version of Asymetrix's *Toolbook*
TBKBASE.DLL	353936	DLL required by Asymetrix's *Toolbook*
TBKCOMP.DLL	105104	DLL required by Asymetrix's *Toolbook*
TBKUTIL.DLL	59008	
HELP.TBK	******	Help file (To be included)

Starting DataBook:

1. Asymetrix's *Toolbook*, and therefore *DataBook*, requires Microsoft's *Windows*. If *Windows* is not currently running on your system, you would start it now by typing ***WIN*** .

2. From the **Program Manager** window, click on the **File** menu and then click on the **Run...** option

3. The dialog box for **Run** is now open, and you type in the **Command Line** as follows: ***C:\TOOLBOOK\TBOOK DATABOOK*** and then click on the **OK** button. The **DataBook** window should now be on the screen.

Using DataBook:

(You can leave *DataBook* at any time by clicking on the File menu and then clicking on Exit.)

1. With the **DataBook** window on the screen, we will now use the "book" metaphor that underlies the design of Asymetrix's *Toolbook*. *DataBook* is like a "workbook," and your first task is to make your own copy of it to use so that you don't mess up the original. One purpose of the first page of *DataBook* is to let you make such a copy.

2. The insertion caret is blinking to the right of **PI:**, which stands for "Principal Investigator." Type in the principal investigator's name in the space provided. It is best *not* to press the **Enter** key here, for it removes the name from the screen (although it does "enter it"). Instead, proceed to the next step.

3. Click on the **Title Book** button. Follow the instructions in the dialog box that appears and then click on the **OK** button. There will be a brief delay while a copy of the book is made. Notice the window title at the top of the screen now contains the name of your book, e.g., **DataBook - MYBOOK.TBK** .

4. Your book has two pages in it, the one you are now on, page one, and page two, the **Data Specification Form** . As with your original copy of *DataBook*, you need to work with one or more copies of the **Data Specification Form** . Again, you don't want to mess up the

original. Click on the **New Page** button. DataBook has now created a copy of the **Data Specification Form** for you to work on and has turned to that page. (You now have three pages in your book.) You must have one copy of the **Data Specification Form** for each independent group in the data set you wish to construct. The name of the principal investigator has been copied onto the form for you, and you are now ready to fill it out.

To put the next set of operations into context, we will create a sample **Data Specification Form** using an example. Assume that you wish to construct the data which would have resulted from the following experiment:

> Twenty college freshman were randomly assigned to treatments, 10 to a treatment group and 10 to a no-treatment control group. The treatment group received a computer-based study skills course during their first semester. Each subject has two measures, high school rank (HSR) and his/her first semester grade point average (GPA). The plan is to carry out an analysis of covariance on GPA using HSR as the covariate.

To get an overview of what is coming, you may want to look at the Appendix. It contains a copy of the filled out Data Specification Form that we are about to enter.

5. Click in the field next to GRP NAME: and type in the group name, Study Skills .

6. Click in the field next to GRP ID #: and type in the group ID number, 1 .

7. Click in the box next to GRP SIZE: and type in 10 .

8. Drop down to the first matrix and click in the first box after LABEL and enter HSR . Move over one box and enter GPA .

9. Move down to the MEAN row and enter 72 and then 2.8 , for the HSR and GPA population means, respectively.

10. Next to SD DEV, enter the population standard deviations, 6 and .3 .

11. INCR refers to the smallest difference between any two values of a variable. HSR is reported in whole numbers, so you enter a 1 . GPA is usually reported to two decimal places; therefore, enter .01 .

12. Next to DIST you can enter the distribution you want to sample. The choices are r for rectangular and n for normal. Since HSR has a rectangular (uniform) distribution, you type in r, and for GPA, you would type in n .

13. MIN and MAX refer to the minimum and maximum values allowed for a variable. For MIN, enter 40 and 1 , and for MAX, enter 99 and 4 . MIN and MAX can be used to truncate distributions. This can introduce, for example, ceiling effects and skewing.

14. Finally, drop down to the correlation matrix (CORR). Notice that a row and column have been automatically labeled with GPA and HSR, respectively, to accommodate the population correlation coefficient between HSR and GPA. Type .5 in the GPA, HSR cell.

15. You are now ready to fill out a Data Specification Form for the control group, and you have two options. You may either create another blank form and fill it in, or make a copy of the one you have just completed. In either case, you need to return to the first page. Click on the Page menu at the top of the window, and then click on First. (You may want to make note of the "keyboard shortcuts" on the Page menu. They speed up moving around the book once you are familiar with the system. Also, you can experiment with using the *Tab* key to move from cell to cell.)

16. In randomized experiments, it is often reasonable to assume that the populations for experimental conditions differ only with respect to their means. We'll assume that here and click on the Copy Page button. A dialog box asks you which page you want to copy. Type in 2 and click on OK, because the Data Specification Form you just completed is page two in the book. A second dialog box asks how many variables you want to copy. For the current example, you would again type in 2 and click on OK. (You may not always want to copy all variables. For example, the Study Skills subjects may have completed an opinion survey on what they liked about the group, something the controls would not have completed.)

17. Repeat steps 5, 6, 7, and 9 to enter values appropriate for the control group.

18. Use the Page menu to return to the first page.

19. Click on Create Data. Messages will begin appearing in the DataBook Response Window. When "Made data for group 2 ." appears in the response window, the data have been created and written to a file in the *TOOLBOOK* directory. The file has the same name as the book with the extension *DAT* (e.g., *MYBOOK.DAT*).

20. You can now click on See File, and the first 20 lines of the file will be listed in the response window. (Note: The columns of data sometimes appear poorly "justified" on the screen. They are correctly justified in the data file itself.)

21. You are now finished, and you can terminate *DataBook* by clicking on the File menu and then clicking on Exit. To avoid losing your work, a dialog box asks you if you want to save your book.

The data file you have created is an ASCII file. Such files can be read by most applications, e.g., word processors, spreadsheets, and statistical packages. You can copy it to a diskette for distribution or to your printer to get a listing.

Finally, this introduction to *DataBook* ends with answers to four questions you might have at this point.

What do I do if I create a page I don't want? If the page isn't on the screen at present, use the Page menu to navigate to it. Click on the Edit menu and then on Select Page. Click on Edit a second time and then on Cut. A dialog box will open, and if you still want to delete the page , click on OK, otherwise click on Cancel.

How do I retrieve a book I've saved so that I can make additions or changes to it? Click on the File menu and then on Open. A dialog box will open with a list of books. Click on the book you want and then on OK.

What is the Repeat Value *button for on the* Data Specification Form*?* You can use it to set a series of fields to the same value. For example, if you wanted seven variables to all have a standard deviation of 10, you would click in the standard deviation cell for the first variable, click on the Repeat Value button, enter 10 in the dialog box that opens and click on OK, then move the "cross" to the cell for the last (seventh) variable and click. This feature is particularly useful when defining large correlation matrices. The first cell you click in defines the upper left corner of a sub matrix, and the second click at the position of the "cross" defines the lower left corner.

What is the purpose of the Repeat Variable *button on the* Data Specification Form*?* The Data Specification Form allows you to define up to 26 variables for each independent group. Sometimes, for example, when you are simulating test items, 26 may not be enough. Repeat Variable allows you to select a variable you have already defined and put a number of randomly generated realizations of that variable in the output file. For example, suppose you have defined three test items which are identical except for their item difficulties, say, .40, .50, .60. If you click on the Label field of the first variable and then click on Repeat Variable, a dialog box opens and asks the number of times you want to repeat the variable in the output file ("repeat" means "repeatedly sample"). Suppose you enter 10 and then click on OK. The variable's label switches to inverse video to indicate that it is "repeated." If you repeat this procedure for the second and third items you have defined, your data file would contain 30 items, or variables.

Appendix

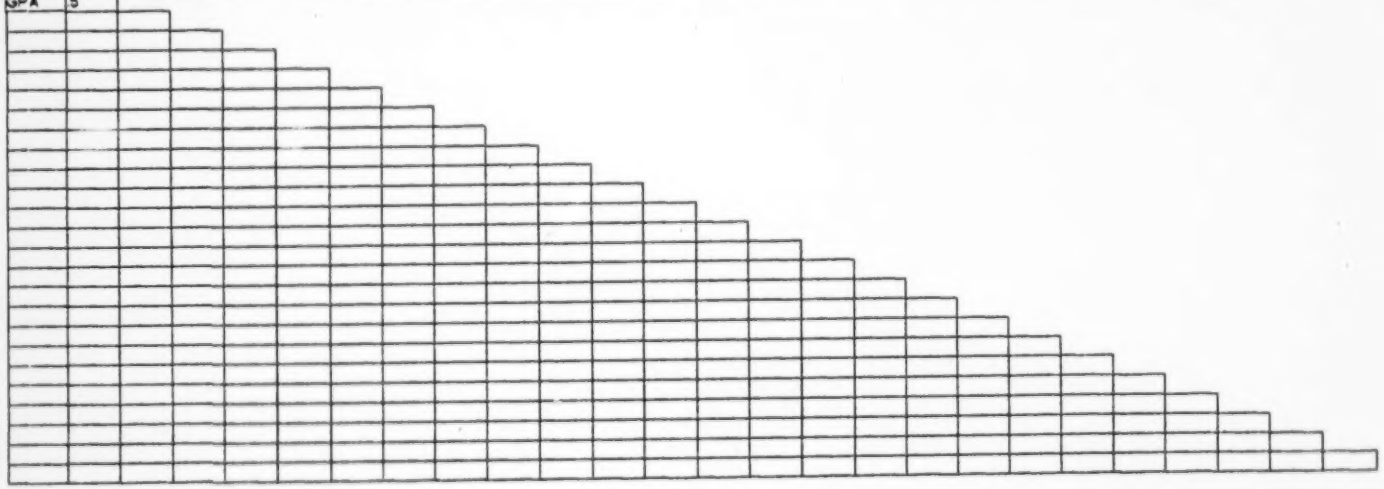
Pt Thomas J. Hummel

GRP NAME: Study Skills GRP ID #: 1 GRP SIZE 10

LABEL	HSR	GPA
MEAN	72	2.8
ST DEV	8	3
INCR	1	.01
DIST	7	h
MIN	40	1
MAX	99	4

Repeat Variable | Repeat Value

CORR	HSR
GPA	.5



Appendix D

Technical Description

There are two components to the *DataBook* program, DATABOOK.TBK and DATABOOK.DLL. DATABOOK.TBK was developed using Asymetrix's *Toolbook* and DATABOOK.DLL was developed using Borland International's *Turbo Pascal for Windows*. Each student has his/her own copy of DATABOOK.TBK which contains one Data Specification Form (DSF) for each independent group of "subjects" in the research design. DATABOOK.TBK handles the interaction with a student as he/she enters his/her DSFs. Once the student has entered the forms, he/she presses the "Create Data" button and DATABOOK.TBK starts the process of data generation. For each DSF, DATABOOK.TBK first checks to make sure the form is complete and then checks for illegal values, e.g., minimum values for variables that are greater than maximums. (DATABOOK.TBK also checks for illegal values at input, e.g., correlation coefficients that are not between -1 and 1.) After DATABOOK.TBK checks a DSF, it calls DATABOOK.DLL and passes the information from the DSF.

DATABOOK.DLL samples N observations from the population defined by the information entered on a single DSF. To do this, DATABOOK.DLL carries out the following operations:

1. The correlations from the DSF are used to construct the p x p population correlation matrix, **P**. The determinant of **P**, $|\mathbf{P}|$, is computed and tested to make sure that $0 < |\mathbf{P}| \leq 1$. This ensures that **P** is positive definite. Students cannot define linearly dependent variables. If these are required, the student must compute them. For example, if a student wanted GRE-V, GRE-Q, and GRE-T = GRE-V + GRE-Q, he/she would could only get GRE-V and GRE-Q from DATABOOK.DLL. GRE-T would be computed after the data set was in hand.

2. Following the methods of Scheuer and Stoller (1962) and Moonan (1957), **P** is factored to produce $\mathbf{C} = [c_{ij}]$, a lower triangular matrix with the property $\mathbf{CC}^T = \mathbf{P}$. If **z** is a p-dimensional vector of independent normally distributed standard scores, i.e., $\mathbf{z} \sim N(\mathbf{0},\mathbf{I})$, then $\mathbf{z}^* = \mathbf{Cz} \sim N(\mathbf{0},\mathbf{P})$.

3. The actual computer algorithm computes the elements of $\mathbf{z}^*$ incrementally, i.e., $z_i^* = \sum_{j=1}^{i} c_{ij} z_j$. *If the distribution for the i^{th} variable was specified as "n", i.e., "normal,"* then z_i^* is transformed to x_i, where $x_i = \sigma_i z_i^* + \mu_i$, and σ_i and μ_i are the standard deviation and mean of the i^{th} variable. Next, w_i, the increment for the i^{th} variable is used to adjust x_i, $x_i = \text{Round}(x_i / w_i) \cdot w_i$, where Round is a function which returns its argument rounded to the nearest whole number. Finally, if x_i is greater than or equal to the minimum specified for i^{th} variable and less than or equal to its maximum, then it is written to the data file. If x_i does not meet this last test, it is discarded and the process is repeated until a satisfactory value is found. *If the distribution for the i^{th} variable was specified as "r", i.e., "rectangular,"* the first step is to compute an upper limit, U_i, and lower limit, L_i, for the distribution of the i^{th} variable. In a uniform distribution, the mean and standard deviation determine the

upper and lower limit and vice versa. Therefore, one could enter a mean and standard deviation on the DSF and minimum and maximum values which would be contradictory. To overcome this, *DataBook* allows zeroes to be entered for the mean and standard deviation of a uniformly distributed variable. In this case, L_i is set to the minimum and U_i to the maximum. If the mean and standard deviation are non-zero, then the entered minimum and maximum are ignored and $L_i = \mu_i - \sigma_i\sqrt{3}$ and $U_i = \mu_i + \sigma_i\sqrt{3}$. The transformation of z_i^* to a uniform distribution is accomplished in the following manner: $x_i = \text{Round}(P(z_i^*)\cdot((U_i - L_i)/w_i + 1) + .499\ldots)$, where $P(z_i^*)$ approximates $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z_i^*} e^{\frac{-z_i^{*2}}{2}}\,dz_i^*$. The approximation $P(z_i^*)$ was obtained from Zelen and Severo (1970, p. 932, 26.2.17). This transformation is similar to converting a normal population of z-scores to percentile ranks. The resulting population of percentile ranks would be uniformly distributed. Since the transformation is not linear, the population correlations involving a z_i^* so transformed will not be equal to the values originally specified in P. However, since the relationship between z_i^* and $P(z_i^*)$ is monotone increasing, i.e., an order preserving transformation, the correlations do not change greatly. (The transformation to uniformity was developed while programming DATABOOK.DLL. No claims are made for its uniqueness or optimality. Tests have shown that it works well in the situation in which it is being used. One might be tempted to generate uniformly distributed variables and then transform them to induce the desired degree of correlation. Multiplying C times a vector of independent uniformly distributed variables would not, in general, work, for as the number of variables increased, the central limit theorem would ensure some elements in the resulting vector would begin to tend toward a normal distribution.) There are other cases were the correlations among the x_i will not be exactly the same as those specified in P. Even for normally distributed variables, the minimum and maximum values can be used to truncate the distribution, and the adjustments made with the increment, w_i, can limit the number of unique values of x_i. Such changes could cause the resulting population correlations to be somewhat different than those specified in P. P is best viewed as a starting point.

4. The procedures described in Step 3 produce one x_i value that is then written to a data file. To give *DataBook* a way to generate more than the 26 variables that can be specified on the DSF, the repetition factor is checked in this step. If the repetition factor for the i^{th} variable, k_i, is greater than one, then Step 3 is repeated an additional k_i - 1 times. Therefore, a total of k_i realizations of x_i will be written to the data file. There is no direct specification of the correlation between the k_i realizations. Whatever correlation they have is induced by the relationship of the i^{th} variable with the preceding variables in the set. For example, suppose there are p = 2 variables, i = 1 to p, and $k_2 = 2$, i.e., there will be two realizations of variable two. In this case, $z_1^* = z_1$ and $z_{2_1}^* = \rho\cdot z_1 + \sqrt{1-\rho^2}\cdot z_{2_1}$, where $z_{2_1}^*$ will be used to get the first realization of the second variable. The second realization of variable two is obtained by generating another standard normal variate, $z_{2_2}^* = \rho\cdot z_1 + \sqrt{1-\rho^2}\cdot z_{2_2}$ (Notice that the value of z_1 is constant

across the two realizations.). Since all variables have unit variance and means equal to zero, the correlation between $z^*_{2_1}$ and $z^*_{2_2}$ will be equal to the expected value of their product, $E(z^*_{2_1} z^*_{2_2}) = E((\rho \cdot z_1 + \sqrt{1-\rho^2} \cdot z_{2_1})(\rho \cdot z_1 + \sqrt{1-\rho^2} \cdot z_{2_2})) = \rho^2$. This result is easily obtained by expanding the product in the expectation on the right side. Note that on taking the expectations of the resulting components of the sum that the expectations of all but one term are equal to zero. This is due to the independence of the normal deviates generated, i.e., $E(z_1 z_{2_1}) = E(z_1 z_{2_2}) = E(z_{2_1} z_{2_2}) = 0$. One application of this feature of *DataBook* is when item responses must be simulated. By setting the maximum of the i^{th} variable to 1, the minimum to 0, and the increment to 1, k_i "ones and zeroes" will be written to the data file. If one wants these item responses to be correlated, then the i^{th} variable ***must be preceded*** by one or more variables that are correlated with it. The correlation among the items will then be a function of the correlations between the preceding variables and the i^{th} variable.

5. At this point, $\sum_{i=1}^{p} k_i$ scores have been written to the data file. ($\sum_{i=1}^{p} k_i = p$ if the variable repetition feature is not used.) This simulates the scores for one subject. The above steps are repeated N times to produce the data for N subjects. DATABOOK.DLL begins a line for each subject with the group identification number followed by the subject number, followed by the $\sum_{i=1}^{p} k_i$ scores. At this point the data file would contain N lines.

6. Having produced a sample of data for one independent group of subjects, DATABOOK.DLL returns control to DATABOOK.TBK. DATABOOK.TBK calls DATABOOK.DLL until all DSFs have been processed.